Spanish Tax Agency

ITS 2.0 implementation experience in HTML5:

www.agenciatributaria.es

MultilingualWeb Workshop
Making the Multilingual Web Work
Rome, 12–13 March 2013
Román Díez González
Spanish Tax Agency

Pedro L. Díez-Orzas
Linguaserve

Linguaserve collaborators:
Giuseppe Deriard-Nolasco, Pablo Nieto Caride, Consuelo Aldana, Félix Fernández
What are we talking about?

1. Introducing the Spanish Tax Agency

2. www.agenciatributaria.es in the MLW-LT project

3. Shifting to HTML5

4. Experience in ITS2.0 annotation:
   a. Automatic annotation of new ITS2.0 metadata
   b. Reusing custom tags for ITS2.0 metadata annotation
   c. Manual ITS2.0 annotation

5. Next steps and some proposals
Spain: General Indicators 2011

Spain is a country regionally structured into 17 autonomous communities and 2 autonomous cities with 5 co-official languages.

- Population: 47,190,493 inhabitants (12.2% foreign residents)

Spanish Tax Agency mission

- Effective application of Spain’s tax and customs structure
- Management of tax resources on behalf of other public administrations when ordered by Law or Agreement

Overall census of obliged taxpayers

- Individual taxpayers: 46,509,231
- Companies: 2,674,547
- Other organisations: 2,293,939

Total taxpayers: 51,477,717
What are we talking about?

1. Introducing the Spanish Tax Agency

2. www.agenciatributaria.es in the MLW-LT project

3. Shifting to HTML5

4. Experience in ITS2.0 annotation:
   a. Automatic annotation of new ITS2.0 metadata
   b. Reusing custom tags for ITS2.0 metadata annotation
   c. Manual ITS2.0 annotation

5. Next steps and some proposals based on experience
(2) The Spanish Tax Agency in MLW-LT


- The MLW-LT Working Group is administered by W3C and receives EC funding (LT-Web) through FP7 in the area of Language Technologies.
(2) The Spanish Tax Agency in MLW-LT

• Online MT System use case components:
  – Multilingual www.agenciatributaria.es (CMS: OpenText WEM)
  – HTML5
  – ITS 2.0
  – Real-time Multilingual Publication System
    • ATLAS (Linguaserve’s Real Time Translation System)
    • Lucy Software MT (Rule-based Machine Translation)
    • MaTrEx from Dublin City University (Statistical Machine Translation)
(2) Online MT System Use Case State

- RTMPS Implementation
  - Prototype 100% (ITS 2.0 definition from Dec 2012)
  - Showcase: preproduction demo (http://its2-aeat.linguaserve.net)
  - ITS 2.0 data categories: 6 (Translate, Localization Note, Language Information, Domain, Provenance, Localization Quality Issue)

- ES-EN total scope: 250 web pages. State:
  - Source language: 30% of target
  - Target language and Post-editing: 30% of target

- ES-FR, ES-DE total scope: 30 web pages. State:
  - Source language: 50% of target
  - Target language and Post-editing: 50% of target

- Testing: pending
(2) Online MT System I18N

Please, see POSTER 4
## (2) MLW-LT Online MT SWOT

### Strengths

- **RTMPS highly reduces:**
  - Translation costs (Quality on-demand)
    - MT + depending on % of post-editing cost reduction increases.
  - Management costs
  - Delivery time
  - Non-invasive technology

### Weaknesses

- **Viability dependent on:**
  - Language combination
  - MT system output
  - Pre-editing and post-editing methodologies and tools (ITS 2.0 and HTML5 compliance)

### Opportunities

- **Profitability:**
  - Websites with more than half a million words
  - Websites with a very high update frequency

### Threats

- **Control, performance and security:**
  - The client might lose control of the translation → user’s control with ITS 2.0
  - Real-time performance
  - Security level
(2) ITS 2.0 benefits for the Spanish Tax Agency

• ITS 2.0 Increases user’s control and automatic decision processes:
  – Translatability and language pair selection (Translate, Language information)
  – Specific terminology to apply (Domain)
  – Activation rules for post-editing (Localization Note)
  – Quality aspects reported to translation consumer or post-editor (Localization Quality Issue)
  – Post-editors judge quality of translation (MT Confidence)*
  – Identification of agents (provenance)
What are we talking about?

1. Introducing the Spanish Tax Agency

2. www.agenciatributaria.es in the MLW-LT project

3. Shifting to HTML5

4. Experience in ITS2.0 annotation:
   a. Automatic annotation of new ITS2.0 metadata
   b. Reusing custom tags for ITS2.0 metadata annotation
   c. Manual ITS2.0 annotation

5. Next steps and some proposals based on experience
(3) Shifting to HTML5: Strategy

- Using ITS 2.0 requires HTML version 5 according to the current W3C specification.
(3) Shifting to shallow HTML5: Modifications

- HTML5 DOCTYPE
- The language page (ISO 639-ISO 3166)
- Self-closed tags not allowed
- Head tags
- Erroneous nesting tags
- Attributes separated by spaces
- Non inclusion of presentation attributes in tags
- Header and body structure needed by tables

- HTML entities instead of special characters
- URLs cannot contain special characters
- ID attribute cannot contain spaces
- Required attributes (e.g. tag "object" must always have the attributes "data" and "type")
- Assessed attributes (e.g. "rel" attribute of tags "a" and "link" must be one from a closed list)
## (3) Shifting to shallow HTML5: Obsolete attributes

<table>
<thead>
<tr>
<th>Tags</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>input</strong></td>
<td>Removed the alt attribute from any input tag that does not contain the attribute &quot;type = 'image'&quot;</td>
</tr>
<tr>
<td><strong>div</strong></td>
<td>Cannot define a &quot;name&quot; attribute in a &quot;DIV&quot; tag</td>
</tr>
<tr>
<td><strong>a</strong></td>
<td>Not allowed to define the attributes &quot;name&quot; and &quot;title&quot; in tag &quot;a&quot;</td>
</tr>
<tr>
<td><strong>embed and object</strong></td>
<td></td>
</tr>
<tr>
<td><strong>table</strong></td>
<td>Not allowed to define the attributes &quot;summary&quot; and &quot;border&quot; in the &quot;table&quot; tag</td>
</tr>
<tr>
<td><strong>img</strong></td>
<td>Not allowed to define the attributes &quot;name&quot; and &quot;border&quot; in the &quot;img&quot; tag</td>
</tr>
<tr>
<td><strong>option</strong></td>
<td>Cannot define the attribute &quot;name&quot; in the &quot;option&quot; tag.</td>
</tr>
<tr>
<td><strong>param</strong></td>
<td>Not allowed to define the attributes &quot;type&quot; and &quot;valuetype&quot; in the &quot;param&quot; tag</td>
</tr>
<tr>
<td><strong>script</strong></td>
<td>Not allowed to define the attribute &quot;lang&quot; except in &quot;JavaScript&quot;, it being case-insensitive in the tag &quot;script&quot;</td>
</tr>
<tr>
<td><strong>br</strong></td>
<td>Cannot define the attribute &quot;clear&quot; in the &quot;br&quot; tag</td>
</tr>
<tr>
<td><strong>background attribute</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No attribute is used to define the &quot;background&quot; in the tags &quot;body&quot;, &quot;table&quot;, &quot;thead&quot;, &quot;tbody&quot;, &quot;tfoot&quot;, &quot;tr&quot;, &quot;td&quot; and &quot;th&quot;.</td>
</tr>
</tbody>
</table>
What are we talking about?

1. Introducing the Spanish Tax Agency
2. www.agenciatributaria.es in the MLW-LT project
3. Shifting to HTML5

4. Experience in ITS2.0 annotation:
   a. Automatic annotation of new ITS2.0 metadata
   b. Reusing custom tags for ITS2.0 metadata annotation
   c. Manual ITS2.0 annotation

5. Next steps and some proposals based on experience
(4) ITS2.0 annotation experience

- Strategy adopted in order to annotate the content with ITS2.0 in an efficient and pragmatic way, considering the pressure and requirements of a real environment.
(4) Automatic ITS2.0 reuse of custom tags

Custom "no translate" tag already exists in the content and is automatically annotated as ITS 2.0 Translate data category:


*Respecting the behaviour of the previous tag and the precedence rules of ITS:

- Addition of ITS default rules for known translatable attributes:
  - <its:translateRule selector="//h:*/@title" translate="yes"/>
  - <its:translateRule selector="//h:*/@alt" translate="yes"/>
(4) Automatic ITS 2.0 annotation: Domain

1. Extracting relevant domains based on the content.
2. Alignment of the domains with each web page.
3. Use of scripts and regular expressions to annotate the content.
4. Document processing:
   i. The selector points to the html root element, indicating that the domain applies to the whole HTML document (inheritance).
   ii. The `domainPointer` attribute indicates where the domain that applies to the selected content is ("Economy and Trade").
   iii. The `domainMapping` maps the domain "Economy and Trade" to "ECON", which will be sent as an understandable parameter to the MT System.
(4) Manual ITS2.0 annotation: Tool

- Quick and pragmatic approach:
  - New HTML Editor plugin created for the ITS 2.0 manual annotation for open source HTML Editor
  - User-friendly interface for the manual insertion of tags.
(4) ITS 2.0 Manual annotation: Translate

• The author must only select the non-translatable element, click on the insertion icon (T) and click on the annotation type: *No Traducir*.
(4) ITS 2.0 Manual annotation: Localization Notes

- Use of the annotation type **Acotar**: The author inserts the annotation text into the box and the software will automatically create the tag.

- The pull-down menu is used to choose the type of localization note. It can either be **description** (descriptiva) or **alert** (alerta).

<p>La disposición trigésima quinta de la Ley del <span its-loc-note="Stands for 'Impuesto sobre la Renta de las Personas Físicas', use acronym in target language" its-loc-note-type="description">IRPF</span></p>
(4) ITS 2.0 Manual annotation: Localization Quality Issue

- Use of the annotation type **Corregir**: The author chooses a type of issue from a pull-down menu, inserts a comment into the box (Comentario), chooses a severity level between 0 and 100 (Severidad) and an optional link to a reference document (documento de referencia), and the software will automatically create the tag.

Online filing can be done by the interested party or by someone representing them. In both cases, an electronic certificate X.509.V3 issued by the <span its-loc-quality-issue-comment="Has previously been translated as 'Royal Mint'. Please be consistent." its-loc-quality-issue-type="inconsistency" its-loc-quality-issue-severity="70">National Coin and Stamp Factory</span>. 

Rome, 12–13 March 2013  MultilingualWeb Workshop  Spanish Tax Agency, IT department
What are we talking about?

1. Introducing the Spanish Tax Agency

2. www.agenciatributaria.es in the MLW-LT project

3. Shifting to HTML5

4. Experience in ITS2.0 annotation:
   a. Automatic annotation of new ITS2.0 metadata
   b. Reusing custom tags for ITS2.0 metadata annotation
   c. Manual ITS2.0 annotation

5. Next steps and some proposals
(5) Next steps and some proposals

• End of Online Translation System MLW-LT use case – June 2013
• Exploring best practices using ITS 2.0 data categories
• Improving real-time translation and multilingual publishing processing by applying extensions, e.g. Readiness:
  – ITS 2.0 extension data category proposal.
  – Linguaserve is applying Readines in both use cases involved:
    • Applied in CMS-TMS showcase (WP3, poster 3)
    • Applicability in Online Translation system (WP4)
  – It indicates the readiness of a document for submission to L10n processes or provides an estimate of when it will be ready for a particular process.
  – It can be used in expert systems for automatic processing.
(5) Next steps and some proposals

• Training and methodologies
  – Pre-editing: ITS2.0 usage and training kits.

• Specific tools
  – Pre-editing:
    • Full HTML5 compliance and ITS2.0 annotation facilities
    • Writing tools for content quality, and controlled language for post-editing output adaptation.
  – Post-editing:
    • Specific language-dependent and language-independent post-editing rules and functionalities.
    • ITS 2.0 assistance and viewing functions for post-editors.